Elsevier Editorial System(tm) for Journal of Applied Research in Memory and Cognition
Manuscript Draft

Manuscript Number: JARMAC-D-14-00059R2

Title: RT-BASED MEMORY DETECTION: ITEM SALIENCY EFFECTS IN THE SINGLE-PROBE AND THE MULTIPLE-PROBE PROTOCOL

Article Type: Original Article

Keywords: Memory detection, Concealed Information, Polygraph, Lie detection, Deception, Reaction time

Corresponding Author: Dr. BRUNO VERSCHUERE,

Corresponding Author's Institution:

First Author: BRUNO VERSCHUERE

Order of Authors: BRUNO VERSCHUERE; Bennett Kleinberg; Kalliopi Theocharidou

Abstract: RT-based memory detection may provide an efficient means to assess recognition of concealed information. There is, however, considerable heterogeneity in detection rates, and we explored two potential moderators: Item Saliency and Test Protocol. Participants tried to conceal low salient (e.g., favourite colour) and high salient items (e.g., first name) and were tested with either the single-probe protocol or the multiple-probe protocol. Experiment 1 was a laboratory study with knowledgeable individuals only (n = 47). Experiment 2 was an Internet study (n = 283), that also included unknowledgeable individuals. High salient items were better detected than low salient items in the laboratory, but not the Internet study (in which the item saliency manipulation was less successful). The multiple-probe protocol outperformed the single-probe protocol in both studies. We conclude that pronounced differences in item saliency affect the validity of RT-based memory detection, and we recommend the multiple-probe protocol for RT-based memory detection.

Suggested Reviewers:

Opposed Reviewers:

***Highlights**

Highlights (3-5; max 125 characters each)

1. The multiple-probe protocol outperforms the single-probe protocol for RT-based memory detection [87 characters]
2. High salient items may be more easily detected than low salient items (Experiment 1) [73 characters]
3. An independent assessment of item saliency can help to elucidate the impact of item saliency on memory detection [97 characters]
4. RT-based memory detection can be have high diagnostic validity in both offline and online research [86 characters]

RT-BASED MEMORY DETECTION: ITEM SALIENCY EFFECTS IN THE SINGLE-PROBE AND THE MULTIPLE-PROBE PROTOCOL

Bruno Verschuere[1, 2, 3*], Bennett Kleinberg[1] & Kalliopi Theocharidou[4]

Affiliation:

1. Department of Clinical Psychology, University of Amsterdam, The Netherlands
2. Department of Psychology, Ghent University, Ghent, Belgium
3. Department of Clinical Psychological Science, Maastricht University, Maastricht, The Netherlands
4. Amsterdam University College, The Netherlands

* Corresponding author: Bruno Verschuere, Department of Clinical Psychology, University of Amsterdam, Weesperplein 4, 1018 XA, Amsterdam, Netherlands. E-mail: b.j.verschuere@uva.nl. Phone: +3125256799

Abstract

(current: 150 words; max 150 words)

RT-based memory detection may provide an efficient means to assess recognition of concealed information. There is, however, considerable heterogeneity in detection rates, and we explored two potential moderators: Item Saliency and Test Protocol. Participants tried to conceal low salient (e.g., favourite colour) and high salient items (e.g., first name) and were tested with either the single-probe protocol or the multiple-probe protocol. Experiment 1 was a laboratory study with knowledgeable individuals only (n = 47). Experiment 2 was an Internet study (n = 283), that also included unknowledgeable individuals. High salient items were better detected than low salient items in the laboratory, but not the Internet study (in which the item saliency manipulation was less successful). The multiple-probe protocol outperformed the single-probe protocol in both studies. We conclude that pronounced differences in item saliency affect the validity of RT-based memory detection, and we recommend the multiple-probe protocol for RT-based memory detection.

Max 6 KEYWORDS: Memory detection, Concealed Information, Polygraph, Lie detection, Deception, Reaction time

Word Count: 5963 (without abstract and refs)

Introduction

Reaction times (RTs) can be used to reveal information that people may not have conscious access to, or are unwilling to report. The possibility to use RTs for lie detection has intrigued researchers for a long time (Jung, 1910). In the present study, we focus on the possibility to detect recognition of concealed information through RTs. While extreme high accuracy of RT-based memory detection has sometimes been obtained (e.g., Seymour et al., 2000), others found poor detection rates (e.g., Matsuda et al., 2009, 2011; Mertens & Allen, 2008). Iacono (2007) concluded that ´whether reaction time alone may itself lead to highly accurate classification of guilty and innocent test takers remains to be determined (pp. 696)´. To explain heterogeneity in detection rates, we explored two potential moderators in this study: Item Saliency and Test Protocol.

RT-based memory detection

RT-based memory detection originates from the more general approach of memory detection originally used in conjunction with a polygraph (also known as the Concealed Information Test or the Guilty Knowledge Test; Lykken, 1959, for a comprehensive review see Verschuere et al., 2011). In RT-based memory detection, RTs are used to infer whether or not an examinee recognizes critical (e.g., crime) information. Someone accused of stealing a laptop, for instance, could be asked to react as fast as possible to possible stolen items (iPod, laptop, wallet, watch, bracelet). Provided the alternatives are well selected, the naïve examinee can be expected to respond similarly to all items. The actual thief, however, is expected to recognize the stolen object (called the *probe*) and react differently to it than to the *irrelevant* items. To assure processing of all the items, the examinee is typically required to make a dichotomous decision, answering NO to all items, except to a dedicated *target* item that requires a YES response. The targets may not only assure semantic processing of the stimuli, but may further help to increase the probe-irrelevant difference in knowledgeable

individuals by inducing a conflict between the urge to press YES to the probes and the task requirement to press NO to them (Suchotzki et al., 2013). The response to the target is typically discarded from the analyses (but see Noordraven and Verschuere, 2013), which focus on the probe-irrelevant contrast. In sum, differential responding to the probe compared to the irrelevant items is taken as an indication of recognition. Several studies showed that RT-based memory detection was very successful (for a review see Verschuere, Suchotzki, & Debey, 2014). Visu-Petra et al. (2012), for instance, instructed half of their participants to commit exam fraud by stealing a CD with exam questions from a laptop bag (guilty condition). Participants in the innocent condition had no knowledge of the exam fraud. Participants had to press the YES as fast as possible for a set of recently memorized target pictures, and NO for all other pictures (including a picture of the laptop bag that contained the stolen CD). Guilty, but not innocent, participants reacted slower to crime-related pictures, and the RT-CIT allowed discriminating guilty from innocent participants with near perfect accuracy. Such studies point to the potential of RTs for memory detection. At the same time, there is reason for caution, as some (ERP) studies found poor detection rates for RTs (e.g., Matsuda et al., 2009, 2011; Mertens & Allen, 2008). The present study sought to explain this heterogeneity by experimentally examining two possible moderators: Test protocol and Item Saliency.

Test protocol: Single-probe versus multiple-probe protocol

The test protocols that have been used in RT-based memory detection differ in several ways. One such aspect is the use of the single-probe versus the multiple-probe protocol (Rosenfeld et al., 2006). In the multiple-probe protocol, items from all categories are all presented completely intermixed in the same block. In the CIT by Visu-Petra described above, the participant may have been presented with a picture of a laptop on one trial, and a picture of a CD on the next trial. In the single-probe protocol, each block is used to test

recognition of a single piece of information. Thus, for instance, first presenting all CDs to examine recognition of the stolen CD, and subsequently present all laptop bags in the next block.

Although RT studies have almost exclusively relied upon the multiple-probe protocol (for a review see Verschuere, Suchotzki, & Debey, 2014), the single-probe protocol is often used by researchers with a prime interest in other (e.g., neural) measures. A disadvantage of the multiple-probe protocol is that the examinee immediately encounters all stimuli, preventing the use of details that reveal the crime-relatedness of other details. In the single-probe protocol one could, for instance, first test on the type of vehicle used in a terrorist attack [Was the bomb in a…train?…plain?…car?…Etc], and subsequently on the type of car [is it a Mercedes? …a Peugeot?...a Suzuki?...Etc], see Meijer, Bente, Ben-Shakhar, & Schumacher, 2013). Both questions cannot be used together in the multiple-probe protocol, because confrontation with the different brands of cars reveals the correct answer with regard to the type of vehicle. The single-probe protocol has been advocated by some as the preferred protocol for the ERP-CIT (Rosenfeld et al., 2006). Specifically, it was argued that the increased complexity of the multiple-probe protocol may *reduce* attention to the probes, and thereby reduce the probe-irrelevant difference. The authors, however, also acknowledged the alternative possibility that the multiple-probe protocol would bring about more attention to the stimuli, and by assuring encoding of the stimuli, would *increase* the probe-irrelevant difference. There is only one study that contrasted the single-probe protocol with the multiple-probe protocol. While overall RTs were higher in the multiple-probe protocol than in the single-probe protocol, Rosenfeld and colleagues (2006) did not find differences in RT detection efficiency between the two protocols. There was a non-significant trend towards better detection with the P300 event-related potential in the single-probe protocol than in the multiple-probe protocol. This study was not conclusive, however, because (1) participants

were not randomly assigned to protocols, (2) the protocol comparison was confounded by differences in the amount and type of stimuli, and (3) the relatively small sample size (n = 9-13 in each cell). The present study re-examined whether and how test protocol affected the validity of RT-based memory detection. Because highly salient items may stand out and grab attention, even in suboptimal protocols, we also included Item Saliency as a possible moderator.

Item Saliency

Item saliency has long been reasoned to be an important factor in memory detection. Lieblich et al. (1976) were among the first to empirically examine the role of stimulus saliency. These authors tested a group of prisoners on 20 autobiographical details (e.g., favourite cigarette brand) and found that the validity of polygraph-based memory detection was lower than that observed in a group of students. However, when restricting the comparison to what they deemed to be the 5 most salient items (e.g., name of the examinee and that of close relatives), detection rate in prisoners was as high as that obtained in students. Relatedly, mock crime studies have found that central details of the crime are better remembered and better detected in the CIT than peripheral crime details (Carmel, Dayan, Naveh, Raveh, & Ben-Shakhar, 2003; Gamer & Berti, 2012; Gamer, Kosiol, & Vossel, 2010; Jokinen et al., 2006; Nahari & Ben-Shakhar, 2011; Peth, Vossel, & Gamer, 2012). It is not always clear, however, how the distinction between high versus low salient (central versus peripheral) details was made, and no study so far formally assessed whether the items indeed differed in the presumed moderator – Item Saliency.

The present study

We examined whether Stimulus Saliency and Test Protocol affect the validity of RT-based memory detection. Participants were tested with either the single-probe or the multiple-probe protocol. Stimuli were categorized a priori as having high versus low salience using an

independent criterion. We expected higher validity (accuracy) for high compared to low salient items and for the multiple-probe compared to the single-probe protocol.

Experiment 1

Method

The study was approved by the ethics committee of the Department of Psychology of the University of Amsterdam (2014-CP-3389).

Participants

Forty-seven (33 females; *M* age = 19.96, *SD* = 1.30) Dutch undergraduates of the University of Amsterdam participated for partial fulfillment of course credits. We excluded data from participants who had less than 50% accuracy on any of the three trial types, considering that such a low accuracy provided an indication that the participants failed to understand or to follow the instructions. Data from 1 participant were excluded because of low probe accuracy, and data from another 6 participants were excluded because of low target accuracy (*M* error rate on target items = 29.48%, *SD* = 14.88). The final sample consisted of 40 participants. The single-probe protocol (n = 21; $_{M}$age = 20.09, SD = 1.30; 16 females) and the multiple-probe protocol (n = 19; M_{age} = 20.21, SD = 1.32; 10 females) did not differ significantly in age, $t(38) < 1$, or gender, $\chi^2(1) = 2.43$, $p = .12$.

Apparatus

Stimuli in the RT-based test were presented by a desktop using Inquisit 3 software (2003), which allows recording of RTs with millisecond accuracy.

Procedure

Participants were tested individually. The whole procedure took about 30 minutes.

After providing written informed consent, participants completed an autobiographical form that asked for demographics and personal information that we would use as probes in the subsequent RT-based memory detection test (first name, last name, favorite dish, favorite

color). We also presented participants with a list of possible first names, last names, dishes and colors that we aimed to use as irrelevant items in the subsequent RT-based memory detection test. Participants were asked to erase items in the displayed list that were of personal significance to them. The predetermined irrelevant and target items were replaced by an alternative if the participant had indicated them to be of personal relevance. Participants waited until the experimenter personalized the script of the RT-based memory detection test.

Next, participants learned the target items as if it was their new identity. This learning phase consisting of a 30 s computerized presentation of a card with the 4 target items (first name: ROBIN; last name: MEYER; favorite color: GREEN; favorite dish: LASAGNA). Memory for the 4 target items was assessed by asking the participants to report all target items. The learning and recall phase was presented twice.

After target memorization, participants conducted the RT-based memory detection test (see below). Finally, as a manipulation check they rated the categories used in the RT-based memory detection test (first name, last name, favorite dish and favorite color) on significance along with 10 other categories (e.g., birthday). Following Dindo and Fowles (2008), personal significance was rated using a 9 point Likert scale (1 = not significant at all to 9 = very significant), with the instruction to judge ´how important, relevant or significant the items are to you, irrespective of whether they are positive or negative´. To explore potential differences between the test protocols, participants also rated the difficulty of the task, how much attention they paid to the stimuli, and how strong the personal items grabbed their attention on a 9-point Likert scale (1 = not at all to 9 = very much).

High versus Low Salient Items

Prior to this study, we had asked 28 undergraduates from Ghent University (Belgium) to rate a number of items (first name, last name, hobby, favorite dish, favorite color, birthday) on personal significance using the 9-point rating scale and instructions provided by Dindo and

Fowles (2008). We used first name and last name for the high salient category (M = 7.64; SD = 1.65), and favourite colour and favourite dish for the less salient category (M = 4.00; SD = 1.17), $t(27) = 9.32$, $p < .001$, $d_{within} = 1.84$.

RT-based memory detection test

On each trial, the participant was presented with a single item in the middle of the screen, either a probe (e.g., the participant´s first name), a target or an irrelevant item. Participants were asked to indicate as fast as possible whether they recognized the stimulus. The YES button (left key press) meant recognition of the stimulus, while the NO button (right key press) meant non-recognition. All participants were instructed to hide their true autobiographical information, pressing NO for probes (own identity) and irrelevant items (unknown identity), and YES to targets (newly acquired identity). There was a short delay (varying between 500, 800 and 1000ms) after key press (or a maximum 1500ms after stimulus onset in case the participant did not respond) before the next stimulus appeared.

Prior to actual test block, there were two practice blocks of 12 trials each (2 targets, 2 probes, 8 irrelevant items). In the first practice block, there was no time limit, and items remained on screen until button press. Error feedback was presented during the first practice block only (i.e., a ´WRONG´ message on bottom of the screen for 500ms after behavioral error). In the second practice phase, there was a time limit with a 'TOO SLOW' warning appearing 800ms after stimulus onset for 500ms on top of the screen. There were four test blocks, each having 72 trials (12 targets, 12 probes, 48 irrelevant items), thus totaling 288 trials. The 800ms time limit was used in all test blocks, and there was a self-paced break in between the blocks.

Single-probe protocol versus Multiple-probe protocol

In the single-probe group, all items in one block belonged to the same category. Thus, there was a block with first names, a block with last names, a block with favorite dishes and a block with favorite colors. Order of the blocks was random. In the multiple-probe protocol, each block contained items from all the 4 categories, presented in random order.

Results

Effect sizes for interaction effects were estimated using Cohen's *f*, with values from .10, .25, and .40 representing small, moderate, and large effects, respectively. We calculated *f* using the following formula: $f = \sqrt{[\eta p^2 /(1 - \eta p^2)]}$ (Cohen, 1988). For follow-up contrasts, we use the following annotations of the Cohen´s *d* effect sizes: in accordance with a meta-analysis by Suchotzki et al. (2014), we calculated the effect size for within-subject contrasts as $d_{within} = M(RT_{(probes)} - RT_{(irrelevants)} / \sqrt{(SD_{(probes)}^2 + SD_{(irrelevants)}^2 - 2*r*SD_{(probes)}*SD_{(irrelevants)}}$, where *r* is the correlation between $RT_{(probes)}$ and $RT_{(irrelevants)}$. For between-subjects contrasts, $d_{between} = (M_{RT(Probe\text{-}Irrelevant\ Difference\ knowledgeable)} - M_{RT(Probe\text{-}Irrelevant\ Difference\ naive)}) / \sqrt{(((n_{knowledgeable} - 1)*SD_{(Probe\text{-}Irrelevant\ Difference\ knowledgeable)}^2 + (n_{naive} - 1)*SD_{(Probe\text{-}Irrelevant\ Difference\ naïve)}^2)/n_{knowledgeable} + n_{naive} - 2)}$. The formulae were adopted from Lakens (2013).

Manipulation Check

As predicted, the significance ratings of the high salient items (M = 7.61; SD = 1.33) were higher than that of the less salient items (M = 4.41; SD = 1.91), $t(39) = 9.45$, $p < .001$, $d_{within} = 1.50$.

RT-based memory detection test

Behavioural errors (i.e., pressing NO to targets or YES to probes or irrelevant items) were excluded from RT analyses, as well as any correct RT smaller than 150ms or greater than 800ms (cf Verschuere et al., 2010), see Footnote1.

The 2 (Protocol: Single-probe versus Multiple-probe protocol) x 2 (Stimulus: Probe versus irrelevant) x 2 (Saliency: High salient vs Less salient) mixed ANOVA on RTs (in ms) indicated that all main effects and 2-way interactions were significant, *F*´s >4, but was subsumed under the significant 3-way interaction, $F(1, 38) = 8.59$, $p = .01$, $f = 0.48$, see Table 1.

Table 1. Average RTs (in ms; SD in parentheses) for high and low salient probe and irrelevant items in the single-probe and multiple-probe protocol in Experiment 1

	Single-probe protocol			Multiple-probe protocol		
	Probe	Irrelevant	Cohen´s d_{within}	Probe	Irrelevant	Cohen´s d_{within}
High	445	377	1.76	499	435	2.23
Salient	(46)	(39)		(41)	(35)	
Low	374	377	-0.11	475	443	1.10
Salient	(41)	(35)		(59)	(44)	
Collapsed	408	377	1.29	487	439	2.70
	(38)	(36)		(46)	(38)	

To break down the interaction, we looked at effects of Stimulus and Saliency in each condition using two separate 2 (Stimulus: Probe versus irrelevant) x 2 (Saliency: High salient vs Less salient) repeated measures ANOVAs. In the single-probe protocol, the significant main effects of Stimulus and of Saliency was subsumed under the Stimulus x Saliency interaction, $F(1, 20) = 68.28$, $p < .01$, $f = 1.84$. The probe-irrelevant difference in the single-probe protocol was bigger for high salient than for low salient items, $t(20) = 8.26$, $p < .001$, $d_{within} = 1.80$, with the probe-irrelevant difference being significant for high salient, $t(20) = 8.05$, $p < .001$, $d_{within} = 1.76$, but not low salient items, $t(20) = 0.51$, $p = .61$, $d_{within} = -0.11$. In the multiple-probe protocol, the significant main effect of Stimulus was subsumed under the Stimulus x Saliency interaction, $F(1, 18) = 9.66$, $p = .01$, $f = 0.73$. The probe-irrelevant difference in the multiple-probe protocol was bigger for high salient than for low salient items, $t(18) = 3.11$, $p < .01$, $d_{within} = 0.72$, yet – unlike the single-probe protocol – was

significant for both the high salient, $t(18) = 9.74$, $p < .001$, $d_{within} = 2.23$, and the low salient items, $t(18) = 4.78$, $p < .001$, $d_{within} = 1.10$.

Subjectively experienced difficulty, attention, and probe pop out

There were no significant differences between the conditions in experienced difficulty of the task (single-probe protocol: $M = 6.00$, $SD = 1.65$; multiple-probe protocol: $M = 5.47$, $SD = 1.64$), $t(37) < 1$, $p = .33$, $d_{between} = 0.32$, how much attention they paid to the stimuli (single-probe protocol: $M = 6.70$, $SD = 1.30$; multiple-probe protocol: $M = 7.05$, $SD = 0.85$), $t(37) < 1.1$, $p = .33$, $d_{between} = 0.32$, and how strong the personal items grabbed their attention (single-probe protocol: $M = 6.00$, $SD = 1.56$; multiple-probe protocol: $M = 6.63$, $SD = 1.64$), $t(37) < 1.23$, $p = .23$, $d = 0.40$.

Discussion

The results of Experiment 1 indicate that part of the heterogeneity in the validity of RT-based memory detection can be explained by item saliency and test protocol. A limitation of Experiment 1 is the lack of a control group consisting of unknowledgeable individuals. While such a control condition is not strictly needed to investigate moderation by item saliency and test protocol, it prevents us from providing a comprehensive picture of the diagnostic efficiency of the RT-based memory detection test (i.e., providing an estimate not only of sensitivity but also of specificity). Study 2 served as a conceptual replication of Experiment 1, including an unknowledgeable control group.

Experiment 2

We recently developed an online version of the RT-based memory detection test (´Memory Detection 2.0´; Kleinberg & Verschuere, 2014). Online RT-based memory

detection allows us to efficiently and validly run well-powered RT-based memory detection research. Participants (n = 283) were tested with either the single-probe protocol or the multiple-probe protocol, and presented with autobiographical probe items presented along within irrelevant items (knowledgeable condition cf Experiment 1) or only with irrelevant items (unknowledgeable condition). Item saliency was manipulated within-subjects. Because participants may be reluctant to provide intimate autobiographical information (e.g., first name, last name cf Experiment 1) on the Internet, we selected new items that had been rated as being low salient (i.e., favorite alcoholic drink and favorite ice-cream) or high salient (i.e., country of origin and birthday).

Method

The study was approved by the ethics committee of the Department of Psychology of the University of Amsterdam (2014-CP-3389).

Participants

There were 289 participants in this study. There were no data stored for 6 participants, most likely due to the use of an outdated browser or operating system. Thus, we had data for n = 283 (M_{age} = 24.98 years, SD_{age} = 11.99). We excluded data of those participants who were not of legal age or indicated an invalid age (e.g., 0) and excluded data of those IP addresses that were recorded more than once to ensure that we did not include data of participants who did the experiment several times (except of the IP addresses of the university computers used on Dutch Science Weekend, see below), leaving n = 248. The majority of our sample were Dutch native speakers (87%), female (76%), and had at least completed university education (76%). We excluded all participants with an error rate of 50% or more on any of the three item types, leaving n = 210.

The final sample consisted of 210 participants (M_{age} = 25.53 years, SD_{age} = 11.30) who had been allocated to one of four conditions. The allocation of participants to conditions was completely random, that is each participant had a probability of .25 to be in either condition. Forty-four participants were in the knowledgeable, multiple-probe condition (M_{age} = 25.64 years, SD_{age} = 11.79; 84% female), 50 participants were in the unknowledgeable multiple-probe condition (M_{age} = 26.96 years, SD_{age} = 12.57; 78% female), 53 in the knowledgeable, single-probe condition (M_{age} = 24.04 years, SD_{age} = 10.76; 76% female), and 63 in the unknowledgeable, single-probe condition (M_{age} = 25.59 years, SD_{age} = 10.39; 73% female). The conditions did not differ in gender, $X^2(1) = 1.92$, $p = .59$, or age, $F(3, 206) = 0.57$, $p = .633$, $f = 0.05$.

Procedure

Participants were recruited (1) through several (pop) science websites that posted a link to our study along with a short recruiting text (referring to a scientific study on ´Keeping Secrets´), including a popular magazine of the University of Amsterdam and Amsterdam University of Applied Sciences (http://www.foliaweb.nl/), a national science festival (http://www.beyondbiennale.nl/discovery%20festival/home/), the Dutch Science Weekend (http://www.hetweekendvandewetenschap.nl/ and http://www.popupwetenschapper.nl/2014/). Data were collected from Sep, 2 to Oct, 10, 2014. Participants took the test at their own time, on their own computer. In addition, we also (2) recruited participants on the Dutch Science Weekend (Oct, 4, 2014), where participants could attend scientific lectures and demonstrations, and take part in our test. These participants were tested in a typical laboratory setting (i.e, on desktops, in cubicles).

Upon accessing the link, they agreed to the informed consent and proceeded to a page where they indicated their gender, age, mother tongue and educational level. On the next page

they were asked to provide 4 autobiographical details (the probes) by selecting one option from a drop-down menu. We asked for their favourite alcoholic drink (e.g., Martini, which also included the option ´non-alcoholic drink´), their favourite ice cream (e.g., Raspberry), their birthday (e.g., 14 October) and their country of origin (e.g., Netherlands). Additionally, we asked them to indicate one other relevant ice cream, alcoholic drink, birthday and country other than their own. These answers were used to optimise the stimuli (see next section).

Participants were instructed to hide their own identity and adopt a different, new identity (the targets), which they learned on the next page (e.g., Grappa, 19 May, Bulgaria, Nougat). To proceed in the task, they had to type in the targets correctly on the following page and were sent back to the target-learning page if they did not recall this identity correctly. Next, they received detailed instructions about the test (e.g. which keys to press) and started the first of three practice phases. For each of the practice phases there were criteria to be met in order to proceed. After the third practice run, participants were told to proceed to the full memory detection test, and upon completion, they rated a number of item categories including the 4 categories used in the test on their relevance using a 9-point Likert scale (1 = not relevant at all, 5 = slightly relevant, 9 = absolutely relevant). Finally, all participants received their results, were debriefed, were thanked for participation and exited the task.

Online CIT

The experimental task was programmed in JavaScript/Jquery and can be accessed via this link: http://www.lieresearch.com/?page_id=616. In the online CIT, we adopted the stimuli optimisation framework introduced in Kleinberg and Verschuere (2014). Specifically, out of a set of six items, one was randomly determined to function as target. For knowledgeable individuals, four of the five remaining items were selected to function as irrelevants, and the autobiographical details served as probes. For unknowledgeable

individuals, one irrelevant item was randomly selected as the probe (but was in fact non-autobiographical for that participant), and the remaining items served as irrelevants. All of these items were then subjected to an automated optimisation that tested for overlap between the true autobiographical probes and other significant items the participant provided before. If there was an overlap (e.g., the randomly determined target was 8 November and the participants true birthday was 8 November), the computer-determined items were replaced by non-overlapping items (e.g., 1 August).

In the memory detection test, participants provided speeded responses to the items by pressing the E key for YES or the I key for NO to the question “Do you recognize this word?” The key meaning and the question remained on the screen for the duration of the test. The stimuli appeared in the centre of the screen for 1500ms or until a key was pressed. If the participants’ key response was incorrect (i.e. pressing the YES key for probes or irrelevants, or the NO key for targets), a red WRONG appeared below the stimulus for 200ms. The response deadline was 800ms and if the key response did not occur before this deadline, a red TOO SLOW appeared above the stimulus for 200ms. We recorded the RTs as the difference between the key response and the appearance of the stimulus. In technical terms, we used a system clock-independent timing method in microseconds. The ISI was randomly either 250, 500, or 750ms.

To allow our participants to become acquainted with the speed and demand of the task, we applied the following step-wise practice procedure: in the first practice phase the stimuli did not disappear automatically after 1500ms and we did not include a “too slow” feedback, so that the speed of the task was entirely within the participant’s control. The second practice run differed from the first in that it did contain the 1500ms loop of stimuli, and the third practice phase contained all features of the full test, that is, 1500ms stimulus display time and too slow message. For each practice phase, the participant had to obtain an

error rate below 50%, a mean RT of less than 800ms, and was not permitted to have RTs below 150ms on more than 20% of the trials. The latter was our safeguard against continuous key holding. Only if all criteria were met, the participant could proceed to the next phase, otherwise the respective phase was repeated. Each practice phase comprised 24 trials and the full CIT consisted of 192 trials.

Results

Manipulation check

Although participants judged the high salient test items (M = 5.43, SD = 2.02) to be higher in saliency than the low salient test items (M = 4.28, SD = 1.77), $t(257) = 9.26$, $p < .001$, $d_{within} = 0.58$, the difference was much less pronounced than in Experiment 1.

RT Analysis

We excluded incorrect trials and RTs below 150 or above 800ms. Table 2 and 3 show the mean RTs for each cell of the experimental design. We conducted a mixed 2 x 2 x 2 x 2 ANOVA with Identity knowledge (unknowledgeable vs. knowledgeable) and Protocol (single-probe vs multiple-probe) as the between-subjects factors and Item Saliency (low salient vs. high salient) and Stimulus (probe vs. irrelevant) on RTs in milliseconds. This ANOVA showed that all main effects were significant, as well as the 2-way interactions of Identity knowledge X Stimulus and of Protocol X Stimulus. These effects were subsumed under the significant three-way interaction between Identity knowledge, Protocol, and Stimulus, $F(1, 206) = 10.45$, $p = .001$, $f = 0.23$. There were no main or interaction effects of Item Saliency, all F´s<2, ps > .16. As can be seen in Table2 (knowledgeable condition) and Table3 (unknowledgeable condition), the 3-way interaction indicates that the multiple-probe protocol outperformed the single-probe protocol, with a greater probe-irrelevant difference in

the multiple-probe protocol than in the single-probe protocol in knowledgeable individuals, $t(95) = 5.15$, $p < .001$, $d_{between} = 1.05$, but not unknowledgeable individuals, $p > .13$. While the multiple-probe protocol outperformed the single-probe protocol, both allowed differentiating knowledgeable from unknowledgeable participants. The probe-irrelevant difference for knowledgeable participants was larger than for unknowledgeable participants for both the multiple-probe protocol, $t(92) = 7.55$, $p < .001$, $d_{between} = 1.56$, and the single-probe protocol, $t(114) = 3.92$, $p < .001$, $d_{between} = 0.73$.

Table 2. Average RTs (in ms; SD in parentheses) for high and low salient probe and irrelevant items in the single-probe and multiple-probe protocol in knowledgeable individuals in Experiment 2

	Single-probe protocol			Multiple-probe protocol		
	Probe	Irrelevant	Cohen´s d_{within}	Probe	Irrelevant	Cohen´s d_{within}
High Salient	435 (59)	418 (44)	0.46	545 (60)	497 (48)	1.29
Low Salient	442 (59)	425 (41)	0.44	541 (58)	493 (50)	1.08
Collapsed	438 (53)	421 (39)	0.59	543 (54)	495 (48)	1.52

Table 3. Average RTs (in ms; SD in parentheses) for high and low salient probe and irrelevant items in the single-probe and multiple-probe protocol in unknowledgeable individuals in Experiment 2

	Single-probe protocol			Multiple-probe protocol		
	Probe	Irrelevant	Cohen´s d_{within}	Probe	Irrelevant	Cohen´s d_{within}
High Salient	427 (46)	426 (40)	0.02	489 (51)	489 (47)	-0.01
Low Salient	426 (54)	431 (44)	-0.12	497 (48)	487 (43)	0.33
Collapsed	426 (42)	428 (39)	-0.09	493 (46)	488 (44)	0.20

Individual Classification

In order to compare the diagnostic efficiency of the RT-based memory detection for low versus high salient test items under the different protocols, we ran Receiver Operating Characteristics analyses (ROC). Following Noordraven and Verschuere's (2013), we calculated a standardized probe-irrelevant difference for each participant using the formula $d_{CIT} = (M_{RT(probes)} - M_{RT(irrelevants)}) / SD_{RT(irrelevants)}$, and examined to what extent this criterion allowed classifying individuals as knowledgeable versus unknowledgeable. The ROC analysis plots sensitivity against the false positive rate across all possible cut-off points. The corresponding area under the curve (AUC) provides an index of diagnostic efficiency with an AUC value of .5 indicating that the test performs at chance level. Values above .5 are indicative of diagnostic power above chance level, with 1 indicating perfect performance. We used the *pROC* R-package for the ROC analyses (Robin et al., 2011).

Table4. Diagnostic efficiency of RT-based memory detection for high and low salient items in the single-probe and multiple-probe protocol in Experiment 2

	Single-probe protocol		Multiple-probe protocol		Difference between protocols*
	ROC (95%CI)	Cohen´s $d_{between}$	ROC (95%CI)	Cohen´s $d_{between}$	
High Salient	.61 (.51 - .71)	0.45	.81 (.72 - .90)	1.32	.004
Low Salient	.65 (.55 - .75)	0.58	.74 (.64 - .84)	1.00	.20
Collapsed	.69 (.59 - .78)	0.72	.86 (.79 - .94)	1.53	.006

Note. * Using DeLong's test for two ROC curves (Robin et al., 2011)

The AUCs are displayed in Table 4. The AUCs for low versus high salient test items did not differ for either protocol, *p*s > .05. Using DeLong's test for two ROC curves (Robin et al., 2011), the multiple-probe protocol was significantly better than the single-probe protocol for the high salient items, and for the high and low items collapsed.

General Discussion

RT-based memory detection appears an efficient means to assess recognition of concealed information. While several studies found extremely high accuracy, others found that RTs could not detect concealed information (for a review see Verschuere et al., 2014). In the present study we investigated the role of two possible moderators: Item Saliency and Test protocol.

Item Saliency

Experiment 1 confirmed the predicted role of item saliency, as the use of high salient items resulted in higher validity than low salient items in both test protocols. Experiment 2 did not replicate the effect of Item Saliency. Although the two studies differ in several aspects (e.g., lab versus online), we think that the most likely explanation is that the item saliency manipulation in Experiment 2 was less successful than in Experiment 1. In Experiment 1, the items clearly differed in judged saliency: More than 3 points on the 9-point scale, representing a very large effect. In Experiment 2, the difference was much less pronounced: Only 1.15 points, representing a moderate effect. The data indicate that pronounced differences in item saliency affect the validity of RT-based memory detection and thereby extend item saliency effects from physiological measures (Carmel, Dayan, Naveh, Raveh, & Ben-Shakhar, 2003; Gamer & Berti, 2012; Gamer, Kosiol, & Vossel, 2010; Jokinen et al., 2006; Lieblich et al., 1976; Nahari & Ben-Shakhar, 2011; Peth, Vossel, & Gamer, 2012) to RTs. We think that the

use of an independent assessment of item saliency will be of great use in future research. It can help to objectify to what extent items differ in saliency, and in the present study helped to clarify why item saliency did not impact upon memory detection efficiency in Experiment 2. Also, it may be worthwhile to extend the assessment of item saliency (1) and include ratings of other possible moderators (e.g., item familiarity), to determine whether the observed differences can be attributed solely to saliency differences, and (2) to expert-ratings of item saliency, allowing to examine – particularly in mock crime research – whether saliency judgements of examinees correspond with those of the examiners.

Test Protocol

The multiple-probe protocol clearly outperformed the single-probe protocol. Across the two studies, the multiple-probe protocol consistently led to very large effects (*d*s > 1) within the knowledgeable individuals - even under suboptimal circumstances (i.e., the use of low salient items). The effects obtained within the knowledgeable individuals with the single-probe protocol were weaker and less stable, and varied from non-significant to very large. With the inclusion of an unknowledgeable control group, Experiment 2 further showed that the diagnostic efficiency of the multiple-probe protocol was better than that of the single-probe protocol. By identifying an important moderator of RT-based memory detection, our findings may help to explain why studies using the single-probe (e.g., Matsuda et al., 2009, 2011; Meijer et al., 2007, Experiment 2; Mertens & Allen, 2008) protocol found poor detection rates for RTs.

One reason for the higher diagnostic efficiency of the multiple-probe protocol is that it is more difficult, as evidenced by the higher overall RTs in the multiple-probe protocol compared to the single-probe protocol (note that RTs were higher not only for probe, but also for irrelevant items). As such, this finding seems to fit with the general *cognitive load hypothesis*, which holds that lie detection is more efficient under high load than under low

load (Vrij et al., 2006). Alternatively, we think that the multiple-probe protocol ensures better processing of the stimuli. Suchotzki and colleagues (2013) showed that the efficacy of RT-based lie detection tasks critically depends on the extent to which they promote processing of the relevant stimulus features (e.g., their truth value). Applied to RT-based memory detection, it is clear that the examiner must try to assure that the examinee discriminates the probes from the irrelevant items whereas the explicit task for the examinee is a mere target versus non-target discrimination. We think that the single-probe protocol allows the examinee to focus upon the target versus non-target dimension, whereby effectively neglecting the probe-irrelevant difference. The subjective ratings in Experiment 1 did not support either the cognitive load or the *relevant feature hypothesis*, so the reasons for the differences between the two test protocols remain to be tested. Because the single-probe protocol has its benefits (e.g., sequential testing), one may also search for ways to assure processing of the probe-irrelevant difference, for instance through stimulus degradation (e.g., %B%R%U%N%O% instead of BRUNO; De Houwer et al., 2001).

Practical Application

The present study has straightforward implications for the applied usage of RT-based memory detection. We recommend the use of the multiple-probe protocol when RTs are the prime measure of interest. Our findings also provide partial support for the recommendation to use high salient items (Osugi, 2011). Provided our findings generalize to crime details, the use of high salient items may increase detection efficiency, particularly under more realistic circumstances (e.g., when there is a delay between crime and test).

Conflict of interest statement

The authors declare that they have no conflict of interest

Footnote1

The original data of both studies are publically available on the Open Science Framework: https://osf.io/kgum2/

References

Carmel, D., Dayan, E., Naveh, A., Raveh, O., & Ben-Shakhar, G. (2003). Estimating the validity of the guilty knowledge test from simulated experiments: The external validity of mock crime studies. *Journal of Experimental Psychology: Applied, 9*, 261-269. http://dx.doi.org/10.1037/1076-898x.9.4.261

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences*. Hillsdale: Lawrence Erlbaum. http://dx.doi.org/10.4324/9780203771587

De Houwer, J., Hermans, D., & Spruyt, A. (2001). Affective priming of pronunciation responses: Effects of target degradation. *Journal of Experimental Social Psychology, 37*, 85-91.http://dx.doi.org/10.1006/jesp.2000.1437

Dindo, L., & Fowles, D. C. (2008). The Skin Conductance Orienting Response to Semantic Stimuli: Significance Can Be Independent of Arousal. *Psychophysiology, 45*, 111-118. http://dx.doi.org/10.1111/j.1469-8986.2007.00604.x

Gamer, M., & Berti, S. (2012). P300 amplitudes in the Concealed Information Test are less affected by depth of processing than electrodermal responses. *Frontiers in Human Neuroscience, 6*:308. http://dx.doi.org/10.3389/fnhum.2012.00308

Gamer, M., Kosiol, D., & Vossel, G. (2010). Strength of memory encoding affects physiological responses in the Guilty Actions Test. *Biological Psychology, 83*, 101-107. http://dx.doi.org/10.1016/j.biopsycho.2009.11.005

Iacono, W.G. (2007). Detection of Deception.. In J.T. Cacioppo, L.G. Tassinary, & G.G. Berntson (Eds.). *Handbook of Psychophysiology* (3rd edition; pp. 688-703). Cambridge, UK: Cambridge University Press.

Inquisit 3.04 [Computer software]. (2010). Seattle, WA: Millisecond Software

Jokinen, A., Santtila, P., Ravaja, N., & Puttonen, S. (2006). Salience of guilty knowledge test items affects accuracy in realistic mock crimes. *International Journal of Psychophysiology, 62*, 175–184. http://dx.doi.org/10.1016/j.ijpsycho.2006.04.004

Jung, C. G. (1910). The association method. *American Journal of Psychology, 21*(April), 219-269. http://dx.doi.org/10.2307/1413002

Kleinberg, B., & Verschuere, B. (2014). Memory detection 2.0: The first web-based memory detection test. University of Amsterdam: Unpublished manuscript.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology, 4*, Article 863, doi: 10.3389/fpsyg.2013.00863

Lieblich, I., Ben-Shakhar, G., & Kugelmass, S. (1976). Validity of the guilty knowledge technique in a prisoner's sample. *Journal of Applied Psychology, 61*, 89-93. http://dx.doi.org/10.1037/0021-9010.61.1.89

Lykken, D. T. (1959). The GSR in the detection of guilt. *Journal of Applied Psychology, 43*, 385–388. http://dx.doi.org/10.1037/h0046060

Matsuda, I., Nittono, H., Hirota, A., Ogawa, T., & Takasawa, N. (2009). Event-related brain potentials during the standard autonomic-based concealed information test. *International Journal of Psychophysiology, 74*, 58-68. http://dx.doi.org/10.1016/j.ijpsycho.2009.07.004

Matsuda, I., Nittono, H., & Ogawa, T. (2011). Event-related potentials increase the discrimination performance of the autonomic-based concealed information test. *Psychophysiology, 48*, 1701-1710. http://dx.doi.org/10.1111/j.1469-8986.2011.01266.x

Meijer, E. H., Bente, G., Ben-Shakhar, G., & Schumacher, A. (2013). Detecting Concealed Information from Groups Using a Dynamic Questioning Approach: Simultaneous

Skin Conductance Measurement and Immediate Feedback. *Frontiers in Psychology, 4,* 1-6. http://dx.doi.org/10.3389/fpsyg.2013.00068

Mertens, R., & Allen J.J. (2008). The role of psychophysiology in forensic assessments: deception detection, ERPs, and virtual reality mock crime scenarios. *Psychophysiology, 45,* 286-98. http://dx.doi.org/10.1111/j.1469-8986.2007.00615.x

Nahari, G. & Ben-Shakhar, G. (2011). Psychophysiological and behavioral measures for detecting concealed information: The role of memory for crime details. *Psychophysiology, 48,* 733-875. http://dx.doi.org/10.1111/j.1469-8986.2010.01148.x

Noordraven., E., & Verschuere, B. (2013). Predicting the sensitivity of the Reaction Time-based Concealed Information Test. *Applied Cognitive Psychology, 27,* 328-335. http://dx.doi.org/10.1002/acp.2910

Osugi, A. (2011). *Daily application of the concealed information test in Japan.* In B. Verschuere, G. Ben-Shakhar, & E. H. Meijer, (Eds.), Memory Detection: Theory and application of the concealed information test (pp 253-275). Cambridge, UK: Cambridge University press. http://dx.doi.org/10.1017/cbo9780511975196.015

Peth, J., Vossel, G., & Gamer, M. (2012). Emotional arousal modulates the encoding of crime related details and corresponding physiological responses in the Concealed Information Test. *Psychophysiology, 49,* 381-390. http://dx.doi.org/10.1111/j.1469-8986.2011.01313.x

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics, 12,* p. 77. DOI: 10.1186/1471-2105-12-77.

Rosenfeld, J. P., Shue, E., & Singer, E. (2006). Single versus multiple-probe blocks of P300-based concealed information tests for self-referring versus incidentally obtained

information. *Biological Psychology, 74*, 396-404. doi:10.1016/j.biopsycho.2006.10.002

Seymour, T.L., Seifert, C.M., Shafto, M.G., & Mosmann, A.L. (2000). Using response time measures to assess "guilty knowledge". *Journal of Applied Psychology, 85*, 30-37. http://dx.doi.org/10.1037/0021-9010.85.1.30

Suchotzki, K., Verschuere, B., Crombez, G., & De Houwer, J. (2013). Reaction Time Measures in Deception Research: Comparing the Effects of Irrelevant and Relevant Stimulus-Response Compatibility. *Acta Psychologica, 144*, 224-231. http://dx.doi.org/10.1016/j.actpsy.2013.06.014.

Suchotzki, K., Verschuere, B., Van Bockstaele, B., Ben-Shakhar, G., & Crombez, G. (2014). *Can reaction time measures differentiate between truthful and deceptive responses? A meta-analysis.* Unpublished manuscript. Ghent, Belgium: Ghent University.

Verschuere, B., Crombez, G., Degrootte, T., & Rosseel, Y. (2010). Detecting concealed information with reaction times: Validity and comparison with the polygraph. *Applied Cognitive Psychology, 24*, 991-1002. http://dx.doi.org/10.1002/acp.1601

Verschuere, B., Ben-Shakhar, G., & Meijer, E. H. (2011). *Memory Detection; Theory and Application of the Concealed Information Test*. Cambridge, UK: Cambridge University Press. http://dx.doi.org/10.1017/cbo9780511975196

Verschuere, B., Suchotzki, K., & Debey, E. (2014). *Detecting deception through reaction times*. In P. A. Granhag, A. Vrij, and B. Verschuere, Deception detection: Current challenges and new approaches. Oxford, UK: John Wiley & Sons, Inc.

Visu-Petra, G., Miclea, M., & Visu-Petra, L. (2012). Reaction time-based detection of concealed information in relation to individual differences in executive functioning. *Applied Cognitive Psychology, 26*, 342-351. doi: 10.1002/Acp.1827

Visu-Petra, G., Varga, M., Miclea, M., & Visu-Petra, L. (2013). When interference helps: Increasing executive load to facilitate deception detection in the concealed information test. *Frontiers in Cognitive Psychology, 4,* 146. http://dx.doi.org/10.3389/fpsyg.2013.00146

Vrij, A., Fisher, R. E., Mann, S., & Leal, S. (2006). Detecting deception by manipulating cognitive load. *Trends in Cognitive Sciences, 10,* 141–142. http://dx.doi.org/10.1016/j.tics.2006.02.003

Table

Table 1. Average RTs (in ms; SD in parentheses) for high and low salient probe and irrelevant items in the single-probe and multiple-probe protocol in Experiment 1

	Single-probe protocol			Multiple-probe protocol		
	Probe	Irrelevant	Cohen´s d_{within}	Probe	Irrelevant	Cohen´s d_{within}
High Salient	445	377	1.76	499	435	2.23
	(46)	(39)		(41)	(35)	
Low Salient	374	377	-0.11	475	443	1.10
	(41)	(35)		(59)	(44)	
Collapsed	408	377	1.29	487	439	2.70
	(38)	(36)		(46)	(38)	

Table 2. Average RTs (in ms; SD in parentheses) for high and low salient probe and irrelevant items in the single-probe and multiple-probe protocol in knowledgeable individuals in Experiment 2

	Single-probe protocol			Multiple-probe protocol		
	Probe	Irrelevant	Cohen´s d_{within}	Probe	Irrelevant	Cohen´s d_{within}
High Salient	435	418	0.46	545	497	1.29
	(59)	(44)		(60)	(48)	
Low Salient	442	425	0.44	541	493	1.08
	(59)	(41)		(58)	(50)	
Collapsed	438	421	0.59	543	495	1.52
	(53)	(39)		(54)	(48)	

Table 3. Average RTs (in ms; SD in parentheses) for high and low salient probe and irrelevant items in the single-probe and multiple-probe protocol in unknowledgeable individuals in Experiment 2

	Single-probe protocol			Multiple-probe protocol		
	Probe	Irrelevant	Cohen´s d_{within}	Probe	Irrelevant	Cohen´s d_{within}
High Salient	427	426	0.02	489	489	-0.01
	(46)	(40)		(51)	(47)	
Low Salient	426	431	-0.12	497	487	0.33
	(54)	(44)		(48)	(43)	
Collapsed	426	428	-0.09	493	488	0.20
	(42)	(39)		(46)	(44)	

Table4. Diagnostic efficiency of RT-based memory detection for high and low salient items in the single-probe and multiple-probe protocol in Experiment 2

	Single-probe protocol		Multiple-probe protocol		Difference between protocols*
	ROC (95%CI)	Cohen´s $d_{between}$	ROC (95%CI)	Cohen´s $d_{between}$	
High Salient	.61 (.51 - .71)	0.45	.81 (.72 - .90)	1.32	.004
Low Salient	.65 (.55 - .75)	0.58	.74 (.64 - .84)	1.00	.20
Collapsed	.69 (.59 - .78)	0.72	.86 (.79 - .94)	1.53	.006

Figure

Footnote

1. Because of the very low error rate (<5%), we do not report error data in full. Also, as most other authors, we excluded targets from the main analyses. For sake of completion, we point out that the target-irrelevant difference predicted the probe-irrelevant difference, $r = .63$, $p < .001$, replicating Noordraven and Verschuere (2013).
2. We ran secondary analyses to examined whether Item Saliency effects were present (1) when only including individuals for whom there was a substantial difference in judged saliency, (2) for idiosyncratic low versus high salient test items. Neither of these analyses showed better detection for high than for low salient items.